IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 26, 2019, accepted July 16, 2019, date of publication July 25, 2019, date of current version August 13, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2931050

Hierarchical Adversarial Network for Human Pose Estimation

IBRAHIM RADWAN[1], (Member, IEEE), NOUR MOUSTAFA[2], (Member, IEEE), BYRON KEATING[3], (Member, IEEE), KIM-KWANG RAYMOND CHOO[4], (Senior Member, IEEE), AND ROLAND GOECKE[5], (Senior Member, IEEE)

[1]Research School of Management, The Australian National University, Canberra, ACT 2601, Australia
[2]School of Engineering and Information Technology, University of New South Wales at ADFA, Canberra, ACT 2610, Australia
[3]QUT Business School, Queensland University of Technology, Brisbane, QLD 4000, Australia
[4]Department of Information Systems and Cyber Security, The University of Texas at San Antonio, San Antonio, TX 78249-0631, USA
[5]Faculty of Science and Technology, University of Canberra, Canberra, ACT 2601, Australia

Corresponding author: Kim-Kwang Raymond Choo (raymond.choo@fulbrightmail.org)

This work was supported in part by the Australian Research Council (ARC) Linkage Project under Grant LP160100910.

ABSTRACT This paper presents a novel adversarial deep neural network to estimate human poses from still images, such as those obtained from CCTV and the Internet-of-Things (IoT) devices. Specifically, the proposed adversarial deep neural network exhibits the spatial hierarchy of human body parts considering the fact that predicting the position of some parts is more challenging than others. The generative and the discriminative portions of the proposed adversarial deep neural network are designed to encode the spatial relationship between the parts in the first stage of the hierarchy (parents) and the parts in the second stage of the hierarchy (children). Each of the generator and the discriminator networks is designed as two components, which are sequentially connected together to infer rich appearance potentials and to encode not only the likelihood of the part's existence but also the relationships between each body part and its parent. The method is evaluated on three different datasets, whose findings suggest that the proposed network achieves comparable results with other competing state-of-the-art approaches.

INDEX TERMS Human pose estimation, hierarchical-aware loss, generative adversarial network, convolutional neural network.

I. INTRODUCTION

Human pose estimation from monocular images, such as those obtained from Internet of Things (IoT) devices, is the process of localizing the landmark positions of the human body parts. Precise human pose estimation is a fundamental step in tasks such as human activity recognition, computational behavior analysis, person re-identification and human-computer interaction. Estimating the 2D pose in unconstrained scenarios is challenging due to the large variations in body poses, the high degree of articulation and hallucination of the body limbs as well as the presence of highly occluded parts.

In the past decade, there have been numerous attempts to develop techniques to overcome the challenges underpinning accurate human pose estimation. The conventional approaches employ hand-crafted features and graphical techniques to generate appearance and pairwise potentials for the body parts. These approaches have played an important role in addressing issues such as variations in clothes, partial occlusion and illumination changes. With the emergence of Deep Convolutional Neural Networks (DCNNs), and the transition from hand-crafted features to DCNNs-based approaches, single [37] and multiple [6] human pose estimation have been impacted by this transition. Many deep networks have been developed so far to tackle other difficulties such as large occlusion and wide range of pose variations, resulting in improved accuracy on pose estimation benchmark datasets [1], [19], [23].

Accurate localization of body parts requires the generation of discriminative features as well as handling the pairwise relationships of the body parts. Approaches have been proposed to achieve this by two subsequent networks (e.g. [5], [38]) or learning a deep network for the appearance features and an explicit kernel for the transformation parameters between the body part positions as in [10].

The associate editor coordinating the review of this manuscript and approving it for publication was Zhaoqing Pan.

FIGURE 1. (a) The hierarchical order of the body landmarks, big dots refer to the anticipated parents; small dots represent their children. The dashed lines refer to the hierarchical relationships between the body parts. (b) Sample results on images of LIP dataset. Despite the occlusion and the high variations in scales and camera views, our approach obtains accurate results.

Moreover, other deep networks have been built to implement this by using only stacked deep networks (e.g. [26]), where the stacked networks and the skip connections between the internal layers help in capturing the structure of the body parts.

The second evolution of DCNN-based pose estimation approaches appeared with the advent of the Generative Adversarial Networks [20], where methods have been developed to employ the novel structure of the deep networks to boost the accuracy of the pose estimation. Chou *et al*. [9] and Chen *et al*. [8] have proposed adversarial networks to learn the appearance features via the generator networks and to encode the structural information of the body parts via discriminator networks. Both approaches predict the positions of all body parts equally, which is not valid as predicting the location of parts with large deformations, such as wrists and ankles, is more complex than predicting less deformable parts, such as the neck or hips. In this paper, we propose a generative adversarial network, which addresses this issue and considers the hierarchical complexities of the body parts. Our proposed networks are inspired by [8], [9], however, the internal structure of the generator and the discriminator is changed to model the hierarchical relationships between the body parts. We also introduce hierarchy-aware terms in the objective functions to regularize the relationships between the parent and the children parts (Fig. 1).

We have tested the proposed networks on three challenging pose estimation benchmarks, where our approach achieves state-of-the-art results on one of these datasets [19] and comparable results to the state-of-the-art methods for the other two datasets [1], [23].

II. RELATED WORK

Human pose estimation retains high attention in the literature. Prior to the dawn of the DCNNs, combinations of hand crafted features, such as histogram of oriented gradients [13], deformable part models [15], [16], [39], and graphical models, such as pictorial structures [2], [14], [17], played a dominant role in estimating human poses from monocular images. These approaches suffered in their capability to produce precise pose estimation in images with high diversity in poses or large occlusions.

With the advent of deep learning and the availability of larger benchmark datasets as well as the existence of devices with high computational power, research in human pose estimation shifted from the classical approaches to DCNN based methods. Toshev and Szegedy [35] introduced the first deep neural network for human pose estimation, called ''DeepPose'', to regress the joint locations directly from the input images. Tompson *et al*. [34] instead combined a DCNN-based model with a graphical approach to infer the spatial relationships between the body parts as well as inferring the joint locations. Bulat and Tzimiropoulos [5] replaced the graphical model of Tompson *et al*. [34] with another deep network to regress the joint locations from the generated heatmaps.

Enforcing the consistency relationships between the joint locations is an effective strategy to accurately infer the position of the body parts. To encode these relationships, Yang *et al*. [38] attached message passing layers on top of the output of the proposed deep network. Hu and Ramanan [21] addressed this problem by using hierarchical rectified Gaussian units, which allow employing the top-down feedback by providing skip connections between the layers of the model. Chu *et al*. [10] introduced transform kernels to encode the relationship between the parts and to use these kernels to infer the locations of the joints dependently. In contrast, Sun *et al*. [32] infer the joint locations by learning the explicit differences between the part pairs to enforce the relationship between these parts. In our proposed method, we follow a similar trajectory and build a model, which learns rich appearance features by building a deep neural network, which enforces the spatial and hierarchical relationships between body parts implicitly. This boosts the performance of the proposed network in predicting the body part locations precisely.

Newell *et al*. [26] introduced an encoder-decoder based network – called *stacked hourglass* – to capture human pose features from an input image at different scales. They used up-sampling layers to decode the appearance features from different scales and allowed skip connections between the convolutional and up-sampling layers, which resulted in encoding the spatial relationships between the parts. They stacked multiple instances of these hourglass based networks and allowed intermediate supervision to infer the human pose from the input image. Their approach achieved significant improvements in human pose estimation. Recently, Chu *et al*. [11] have used stacks of hourglass based networks to generate attention maps, followed by recursively using a Conditional Random Field (CRF) [40] to model the correlations between neighboring attention maps. In our method, the generator of our designed adversarial network follows a

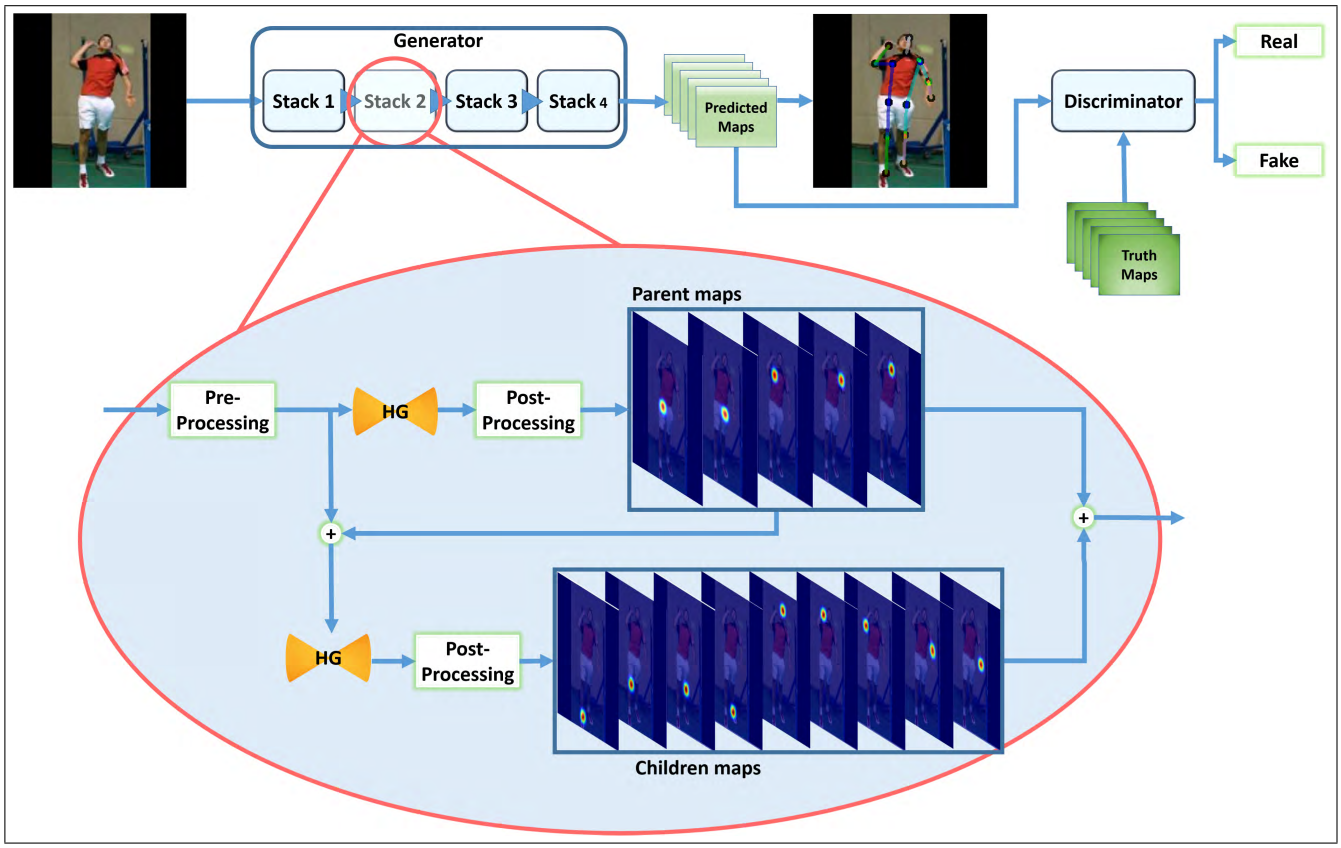


FIGURE 2. Proposed framework: The input image is forwarded through the stacks of the generator and the estimated confidence maps are extracted. The discriminator distinguishes between the real and the estimated maps. The internal steps of each stack are highlighted inside the oval shape.

similar structure by stacking multiple hourglass based networks to capture the appearance features of the body parts. However, we change the design of each stack such that the hierarchical structure of the human body parts is encoded. This helps in enforcing the spatial and hierarchical relationships associated with the body parts as well as enabling end-to-end learning without the need for extra pairwise handling layers.

In [42], hard-keypoints mining was utilized by sorting the estimated keypoints based on their confidence scores and then penalizing the estimator only with the loss of the keypoints, which are estimated with high loss (*i.e.* low scores). This helps the estimator to focus on these hard mining keypoints. Our proposed method provides a general framework for this motivation, where hard-keypoint mining can be seen as a special case of the proposed method. In our method, firstly, we use the cascaded structure for the networks with intermediate supervision. This will lead to making the subsequent networks implicitly look after the keypoints with low scores and refine the entire estimated keypoints. Secondly, the use of an explicit hierarchical structure, by dividing the network into the parents and the children parts, will explicitly force the estimator to focus on the children parts, which mostly will be estimated with lower confidence scores than the parent parts.

Chen *et al.* [8] enforced the prior structure of the human body by proposing a generative adversarial network with two discriminators to penalize the generation of implausible poses. The generator of their model generated confidence maps for the joint locations as well as for the occlusion. Similarly, Chou *et al.* [9] also proposed a similar learning paradigm, such that a generator estimates the confidence maps of the body parts and a discriminator distinguishes between ground-truth confidence maps and the generated ones. In our method, we follow these two approaches and develop a generative adversarial based network. However, we change the design of the underlying generator and discriminator to encode both the structure and hierarchy of the body parts. Recently, Nie *et al.* [27] proposed a hierarchy-aware method to involve the hierarchical structure of the body parts. The hierarchical units of their network work as a post-processing step to fine-tune the predictions of the body parts. In contrast, we design our network to encode the hierarchical structure of the body parts implicitly, which enables our network to be trained end-to-end.

III. PROPOSED METHOD

The proposed model is a generative adversarial network (GAN) with two parts: generator and discriminator. As shown in Fig. 2, the underlying structure of the generator and the discriminator is composed of an *hourglass* based network [26], which is an encoder-decoder network with skip connections and up-sampling layers to capture the diversity

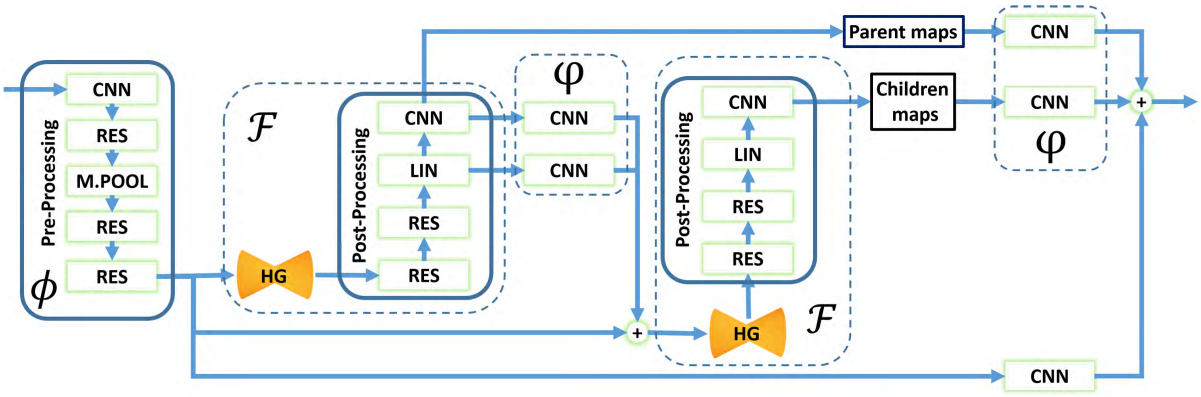


FIGURE 3. The detailed interconnections between the modules in each stack of the proposed network.

in the human pose variations from different scales. The input and output of the hourglass networks are pre-processed and post-processed with residual and linear blocks to prepare the features and to activate valuable confidence score maps for the estimated body part locations.

The RGB images are forwarded through the generator network to estimate the confidence maps of 16 body parts. Each value in a confidence map indicates the likelihood of a specific body part to exist at the pixel coordinates of this value. The discriminator exhibits a similar structure as the generator. It distinguishes, as in the traditional GANs, between the real and generated poses. The discriminator allows the network to implicitly encode the biological structure of the generated poses according to a prior structure.

The articulated structure of the human body results in apparent large deformations of joint locations on the limbs (such as wrists, elbows, knees and ankles), more so than for joints on the torso (such as shoulders and hips), which are less deformable in most cases. Estimating these body parts is challenging and requires a huge and diverse amount of training samples. To address these issues, we modify the structure of the generator and discriminator networks and use dual hourglass networks, which are connected to leverage the spatial hierarchy of the human body parts. Also, we have introduced new loss function terms to regularize the relationship between the parent and children parts. The design of the hierarchical adversarial network (HAN) and the new *hierarchy-aware* loss terms help in accurately estimating the position of body parts, which are largely deformable or highly occluded.

The generator is designed to act as the pose estimator, which employs the hierarchical structure of the body parts to estimate the locations of the body joints hierarchically. However, due to occlusion, lighting conditions, shortening and the high articulation levels, there is still a need to discriminate the good estimations of the whole body pose from the poor-estimated ones. The reasons for using the discriminator network in the proposed method is to: 1) reject the estimations, which are quite unlikely stemming from natural body poses; and 2) to prevent the generator from focusing only on the apparent poses with sufficient appearance features, but also focusing on the poses, which are highly occluded. These two characteristics are incorporated using the auxiliary adversarial loss to boost the performance while training the generator.

A. HIERARCHICAL GENERATIVE NETWORK

The generator is a Fully Convolutional Network (FCN), which infers the confidence maps of the body parts given the input images. The generator is composed of multiple stacks of the dual-hourglass based networks, which are connected sequentially. The interconnections and internal structure of each stack of the dual hourglass are explained in Fig. 3.

The network receives an input image I of size 256×256 pixels. Then, pre-processing steps are performed to extract the pre-mature features using a Convolution Neural Network (CNN), residual blocks and max-pooling modules. This results in down-sampling the size of the extracted features to 64×64. Then, these features form the input to the first hourglass network, which is responsible for estimating the confidence maps of the parent parts. The post-processing step in Fig. 3 consists of residual blocks, linear activation layers and CNN modules, which are used to encode the appearance features extracted from the hourglass network into confidence maps with the same number of parent parts.

In our method, we define the first set of parts to be estimated using the first hourglass network as the parents and those, which are estimated using the second hourglass network, as the children. The choice of the parent parts is derived from the fact that the parts, which are attached to the torso area of the human body are less deformable and can be extracted by a smaller number of layers. These parts are shoulders, hips and neck. Then, confidence maps are processed by different modules and concatenated with the pre-mature features that are extracted early in the network to estimate the locations of the children parts using the second hourglass. The outputs of the dual hourglass networks are concatenated and passed to the next stack in the network. The output of the last stack is the final generated map, which is used to train the discriminator.

The Generator $G : R^M \rightarrow R^N$ maps an input image I into dual sets of confidence maps, $\mathcal{C}_{pa}, \mathcal{C}_{ch}$ for parents and children respectively. Mapping is performed by learning a function $\mathcal{F}_i$ as follows:

$$\mathcal{C}^i_{pa} = \begin{cases} \mathcal{F}_i\big(\phi(I)\big) & \text{if } i = 1 \\ \mathcal{F}_i\big(\phi(I), \varphi(\mathcal{C}^{i-1}_{pa}, \mathcal{C}^{i-1}_{ch})\big) & \text{if } i > 1 \end{cases} \tag{1}$$

$$\mathcal{C}^i_{ch} = \begin{cases} \mathcal{F}_i\big(\phi(I), \varphi(\mathcal{C}^{i}_{pa})\big) & \text{if } i = 1 \\ \mathcal{F}_i\big(\phi(I), \varphi(\mathcal{C}^{i}_{pa}), \varphi(\mathcal{C}^{i-1}_{pa}, \mathcal{C}^{i-1}_{ch})\big) & \text{if } i > 1 \end{cases} \tag{2}$$

where $\phi(I)$ is the function, which applies the pre-processing modules mentioned in Fig. 3 to the input image I, $\varphi(.)$ represents applying CNN modules to the confidence maps of previous steps, and $\mathcal{F}(.)$ is the combination of applying hourglass network and the post-processing steps. In Fig. 3, we label the modules of the network with the corresponding symbols of these functions for illustration.

The confidence maps for both the parent and children parts are then concatenated as the output of the generator for each stack. The predicted poses are inferred using the following equation:

$$\hat{Y} = \underset{x,y}{\text{soft-argmax}}\Big(\sum_{i=1}^{n}(\{\mathcal{C}^i_{pa}, \mathcal{C}^i_{ch}\})\Big) \tag{3}$$

The estimated pose $\hat{Y}$ is computed by summing up the confidence maps through all sequentially connected stacks and selecting the x and y locations using the soft-argmax over the confidence scores of the concatenated maps. The soft-argmax is used instead of the hard-argmax to facilitate computing the gradients in the back propagation process. This makes the whole pipeline trainable end-to-end. We follow [41] to extract the coordinates of the estimated body parts using the soft-argmax on the confidence maps.

The generator network, in itself without the adversarial branch, is trained by minimizing the following loss function:

$$\begin{aligned} \mathcal{L}_{G(I)} = \frac{1}{4nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\Big(&\|Y^j_{pa} - \hat{Y}^j_{pa}\|^2 + \|Y^j_{ch} - \hat{Y}^j_{ch}\|^2 \\ &+ \|D^j_{hier} - \hat{D}^j_{hier}\|^2 + \|\Theta^j_{hier} - \hat{\Theta}^j_{hier}\|^2\Big) \end{aligned} \tag{4}$$

It consists of four terms: the first two terms represent the difference between the estimated confidence maps and the ground truth for both the parent and children parts, respectively. The third and fourth terms represent the *hierarchy-aware loss* terms between the parents and children. These two terms consider the distance D and the angle Θ between the locations with the maximum scores for the parents and children, and compare these differences with the corresponding differences in the ground truth maps. The newly added hierarchy-aware loss terms help in regularizing the large deformation of anticipated locations of the children parts. This constrains the search space for the expected locations of the children parts with respect to the parent parts, which have been estimated at an earlier stage. The loss value of Eq. 4 is averaged by the training samples m and the number of the stacks n.

B. HIERARCHICAL DISCRIMINATIVE NETWORK

The discriminator $D : R^N \rightarrow R^N$ exhibits a similar structure as the generator network with one stack of dual hourglass based network. One stack is used for reconstructing the confidence maps of the parent parts and another one for reconstructing the confidence maps of the children parts. Dividing the structure of discriminator network into two stages lets the discriminator consider the hierarchical structure of the body parts when reconstructing the output maps.

Inspired by [8], the role of the discriminator is to distinguish between fake poses and real poses. Specifically, due to occlusion or large variations in the generated poses, we split the discriminative networks into two parts, D_{pa} and D_{ch}. D_{pa} is responsible for distinguishing the fake poses of the parent parts from the corresponding real poses. D_{ch} achieves the same role for the children parts. The reconstructed maps of the two discriminators are concatenated together to represent the output of the discriminative network.

Training the discriminator is crucial as the discriminator tends to reconstruct from the real poses faster than from the generated poses, which makes the discriminator unable to discriminate the generated poses. To tackle this issue, we followed Berthelot *et al.* [4] by using a balance strategy between reconstructing the real and generated poses. The discriminator is trained with the following loss function:

$$\mathcal{L}_D = \mathcal{L}_{real} - k_t \cdot \mathcal{L}_{fake} \tag{5}$$

where k_t is the balance term, and $\mathcal{L}_{real}$ and $\mathcal{L}_{fake}$ are the loss terms of the real and generated poses, which are computed as follows:

$$\begin{aligned} \mathcal{L}_{real} = \frac{1}{2m}\sum_{j=1}^{m}\Big(&\|C^j_{pa} - D(C^j_{pa}, I)\|^2 \\ &+ \|C^j_{ch} - D(C^j_{ch}, I)\|^2\Big) \end{aligned} \tag{6}$$

$$\begin{aligned} \mathcal{L}_{fake} = \frac{1}{2m}\sum_{j=1}^{m}\Big(&\|\hat{C}^j_{pa} - D(\hat{C}^j_{pa}, I)\|^2 \\ &+ \|\hat{C}^j_{ch} - D(\hat{C}^j_{ch}, I)\|^2\Big) \end{aligned} \tag{7}$$

where $D(., I)$ is the output of the discriminator, which applies to input image I concatenated with the real or generated confidence maps. To control the balance of the running competition between the output of the generator and the output of the discriminator, the balance term $k_t \in [0, 1]$ is updated at each training step t as follows:

$$k_{t+1} = k_t + \lambda_k(\gamma\mathcal{L}_{real} - \mathcal{L}_{fake}) \tag{8}$$

where λ_k is the learning rate of the balance term k, which is the proportional gain of k over time. The diversity term γ plays a critical role in achieving the balance as lower values of γ lead the discriminator to focus more on auto-encoding the real poses, while higher values lead the discriminator to focus more on the diversity of the generated poses. To maintain this balance, we use $\gamma = 0.5$ in the experiments.

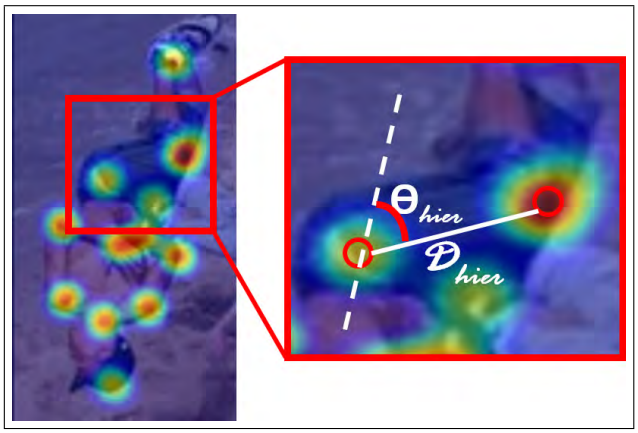


FIGURE 4. Illustration of the hierarchy-aware loss terms. (Left) The confidence maps of the generated poses. (Right) Highlighted example of the distance $\mathcal{D}_{hier}$ and the angle Θ_{hier} of the locations with maximum scores between a parent and a child part.

C. HIERARCHICAL ADVERSARIAL TRAINING

Given M training samples $\{I^j, Y^j\}_{j=1}^{M}$, the two parts of the proposed network are trained simultaneously in a supervised manner with intermediate supervision for the stacks of the generative network. We feed batches of input images through the generative network, which estimates confidence maps for the parent and the children parts. Comparing these confidence score maps with the corresponding ground truth maps, as in Eq. 4. Then, real poses and the input images are fed through the discriminative network to reconstruct the output of the real poses. The parameters of the discriminative network will be updated based on computing the gradients of the loss function, Eq. 5. Then, the parameters of the generative network are updated based on computing the gradients of the following loss function:

$$\mathcal{L}_G = \mathcal{L}_{G(I)} + \lambda_G \mathcal{L}_{fake} \tag{9}$$

where λ_G is a hyper-parameter of the weight of how much emphasis is put on the adversarial loss, which is equivalent to $\mathcal{L}_{fake}$. The training steps are summarized in Alg. 1.

IV. EXPERIMENTS AND RESULTS

In this section, we describe the experiments, which we conduct on three challenging human pose estimation datasets, and their results. (i) The Leeds Sports Poses (LSP) dataset [23] and its extended version, which both contain 12K images, with 11k image for training and 1k for testing. (ii) The MPII dataset [1], which contains 25k images with 40k people performing a wide range of human activities. (iii) The Look Into People (LIP) dataset [19], which contains 50k images with 16 key points annotated for human pose estimation. This dataset has been collected from real-world scenarios with a wide range of poses and views. It contains many images with heavy occlusions and low resolutions.

Algorithm 1 Training Steps of the Hierarchical Adversarial Network (HAN)

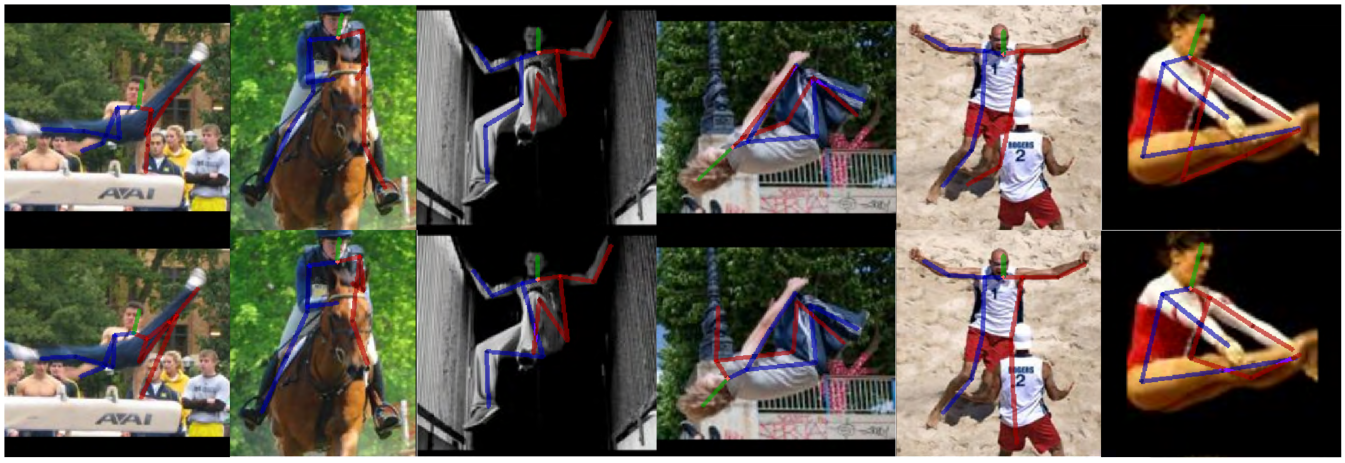
Input: $\{I\}$, $\{\mathcal{C}_{pa}, \mathcal{C}_{ch}\}$
Output: $\hat{Y}$, Eq. 3
while $\hat{Y}$ *still improves for validation set* **do**
  Apply the discriminator network on $\{\mathcal{C}_{pa}, \mathcal{C}_{ch}\}$
  Compute $\mathcal{L}_{real}$ using Eq. 6
  Back propagate the gradients of L_{real} into the Discriminator
  Apply the generator network on the input images
  Compute $\mathcal{L}_{fake}$ using Eq. 7
  Back propagate the gradients of $\mathcal{L}_{fake}$ into the Discriminator
  Update the parameters of discriminator after computing loss using Eq. 5
  Compute $\mathcal{L}_G$ using Eq. 9
  Update the parameters of the generator using gradients of $\mathcal{L}_G$
  Extract $\hat{Y}$ using Eq. 3
end

TABLE 1. *PCK* with $r = 0.2$ on LSP dataset.

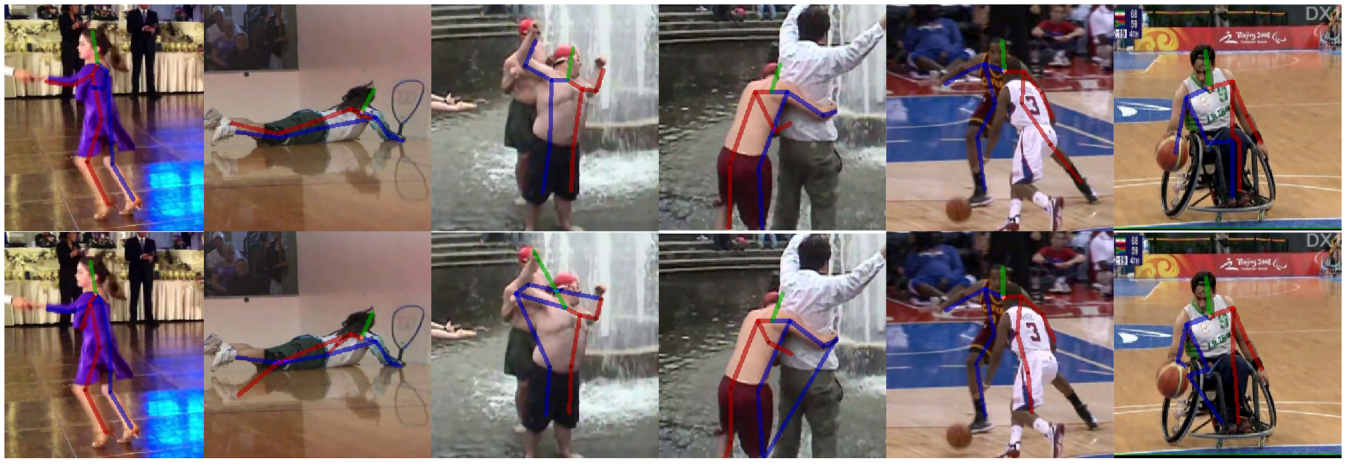
Method	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	Mean
Belagiannis&Zisserman [3]	95.2	89.0	81.5	77.0	83.7	87.0	82.8	85.2
Lifshitz *et al.* [25]	96.8	89.0	82.7	79.1	90.9	86.0	82.5	86.7
Pishchulin *et al.* [29]	97.0	91.0	83.8	78.1	91.0	86.7	82.0	87.1
Insafutdinov *et al.* [22]	97.4	92.7	87.5	84.4	91.5	89.9	87.2	90.1
Pishchulin *et al.* [30]	97.8	92.5	87.0	83.9	91.5	89.9	87.2	90.1
Wei *et al.* [36]	97.8	92.5	87.0	83.9	91.5	90.8	89.9	90.5
Bulat&Tzimiropoulos [5]	97.2	92.1	88.1	85.2	92.2	91.4	88.7	90.7
Chen *et al.* [8]	98.5	94.0	89.8	87.5	93.9	94.1	93.0	93.1
Chou *et al.* [9]	98.2	94.9	92.2	89.5	94.2	95.0	94.1	94.0
Proposed method	**98.5**	**95.1**	**92.4**	**89.9**	**94.9**	**95.4**	**94.6**	**94.4**

A. IMPLEMENTATION DETAILS

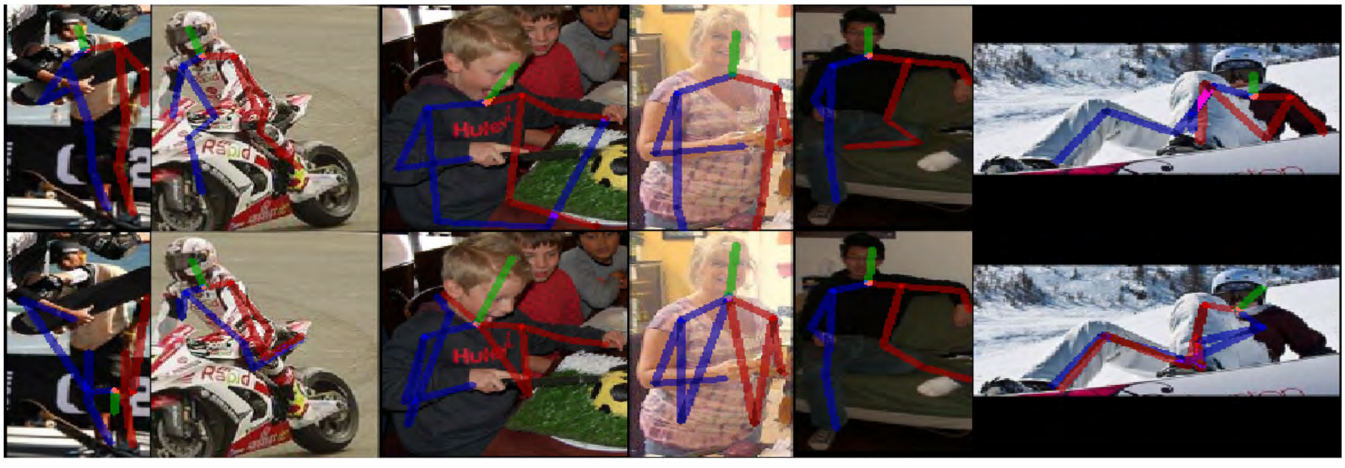
The training data of both MPII and LSP datasets are concatenated and the proposed model is trained for 200 epochs. For the LIP dataset, the model trained on the MPII and LSP datasets is fine-tuned on the training set of this dataset for another 50 epochs. Inspired by [26], [37], the input images are resized and cropped to 256×256 pixels with respect to the annotated body position and scale. Data augmentation is also employed to increase the generalization of the trained models and to prevent over-fitting. Torch7 [12] is used for training and testing. The models are trained on a Titan XP GPU, where the mini-batch size for all of the trained models is 3. We have used RMSProp for optimizing the parameters of the generator and discriminator networks. The learning rate, which is used for the generator network started with 2.5×10^{-4} and then it has been reduced by a factor of 10 at the 101^{th} and the 151^{th} epoch. For the discriminator network, we have used 8×10^{-5} as the learning rate. The balance term k_t is initialized with 0 and has been clipped over iterations with the weight parameter, λ_k as in Eq. 8, where the $\lambda_k = 0.001$ for all of the experiments. The weight of the adversarial loss λ_G in Eq. 9 has been set to 0.0001 in all of the experiments.



(a) LSP dataset samples



(b) MPII dataset samples



(c) LIP dataset samples

FIGURE 5. Qualitative comparison between the proposed method (top) and Chou *et al.* [9] (bottom) for samples from (a) LSP dataset, (b) MPII dataset and (c) LIP dataset. The proposed method shows accurate and coherent poses in difficult conditions.

B. RESULTS AND DISCUSSION

1) LSP DATASET

We follow the previous work of [8], [22], [29] and use the Percentage Correct Keypoints (*PCK*) metric for evaluating the proposed method on the LSP dataset, with 0.2 as the matching tolerance threshold (r).

The results are presented in Table 1. Our model achieves 94.4% and outperforms Chen *et al.* [8] and Chou *et al.* [9],

which are considered as the closest comparable approaches to our proposed method. More specifically, comparing with the results of using an hourglass based network only as in [26] or using a generative adversarial based network, where all of the parts are treated equally as in [8], [9], our model achieves better results for challenging parts such as the elbow, wrist, knee and the ankle body parts. This emphasizes the importance of considering the hierarchical structure of the body parts while estimating the human poses and the validity of the proposed network.

Moreover, we present a qualitative comparison in Fig. 5-(a) between the output of the proposed model and Chou *et al*. [9] on testing samples of LSP dataset. In this figure, we illustrate the ability of the proposed method to obtain better results and the capability to produce coherent and plausible poses via encoding the hierarchical structure of the body parts into the proposed network.

2) MPII DATASET

The MPII dataset represents the most challenging available benchmark for human pose estimation because it contains a wide range of pose variations for many people who perform different activities. In Table 2, we present the PCK^h results on the testing set of the MPII dataset. PCK^h is the same as the normal PCK, except the matching threshold is computed according to the size of the head not the torso. In this table, all of the results are reported with 0.5 as the tolerance threshold. Our results are comparable with the current state-of-the-art results on this dataset. Also, the proposed method has shown better results for most of the body parts compared with [8] and [9] (Table 2).

TABLE 2. PCK^h with $r = 0.5$ on MPII dataset.

Method	Head	Sho.	Elb	Wri.	Hip	Knee	Ank.	Mean
Pishchulin *et al*. [29]	74.3	49.0	40.8	34.1	36.5	34.4	35.2	44.1
Tompson *et al*. [34]	95.8	90.3	80.5	74.3	77.6	69.7	62.8	79.6
Carreira *et al*. [7]	95.7	91.7	81.7	72.4	82.8	73.2	66.4	81.3
Tompson *et al*. [33]	96.1	91.9	83.9	77.8	80.9	72.3	64.8	82.0
Hu&Ramanan [21]	95.0	91.6	83.0	76.6	81.9	74.5	69.5	82.4
Pishchulin *et al*. [30]	94.1	90.2	83.4	77.3	82.6	75.7	68.6	82.4
Lifshitz *et al*. [25]	97.8	93.3	85.7	80.4	85.3	76.6	70.2	85.0
Gkioxary *et al*. [18]	96.2	93.1	86.7	82.1	85.2	81.4	74.1	86.1
Rafi *et al*. [31]	97.2	93.9	86.4	81.3	86.8	80.6	73.4	86.3
Belagiannis&Zisserman [3]	97.7	95.0	88.2	83.0	87.9	82.6	78.4	88.1
Insafutdinov *et al*. [22]	96.8	95.2	89.3	84.4	88.4	83.4	78.0	88.5
Wei *et al*. [36]	97.8	95.0	88.7	84.0	88.4	82.8	79.4	88.5
Bulat&Tzimiropoulos [5]	97.9	95.1	89.9	85.3	89.4	85.7	81.7	89.7
Newell *et al*. [26]	98.2	96.3	91.2	87.1	90.1	87.4	83.6	90.9
Tang *et al*. [43]	97.4	96.4	92.1	87.7	90.2	87.7	84.3	91.2
Chu *et al*. [11]	98.5	96.3	91.9	88.1	90.6	88.0	85.0	91.5
Yang *et al*. [37]	98.5	96.7	92.5	88.7	91.1	88.6	86.0	92.0
Ke *et al*. [24]	98.5	96.8	92.7	88.4	90.6	89.4	86.3	92.1
Tang *et al*. [44]	98.4	96.9	92.6	88.7	91.8	89.4	86.2	92.3
Chou *et al*. [9]	98.2	96.8	92.2	88.0	91.3	89.1	84.9	91.8
Chen *et al*. [8]	98.1	96.5	92.5	88.5	90.2	89.6	86.0	91.9
Proposed method	98.5	96.7	92.5	88.7	91.1	88.6	86.0	92.0

We have highlighted the differences in the achieved results of the proposed method with respect to the previously developed methods in Fig. 6. In this figure, we show the PCK^h values for the average of body parts as well as for wrists and ankles, which are the most challenging parts to be estimated. Our method overcomes the results of the other approaches in the average of all parts as well as in the wrists and ankles parts. This again demonstrates the robustness of the proposed method in estimating difficult body parts. This also is shown in Fig. 5-(b), which depicts the results of our proposed method against the results of [9] on samples from the MPII dataset.

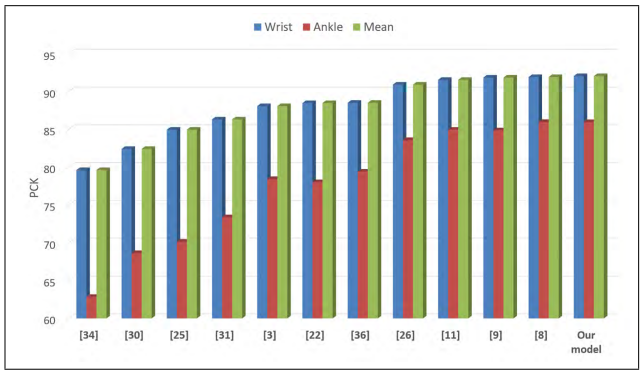


FIGURE 6. The PCK^h values (Y-axis) of the proposed method versus the PCK^h values of other approaches (X-axis) for wrist and ankle parts as well as for the mean of all of the body parts.

3) LIP DATASET

Table 3 presents the PCK^h results of the proposed method compared with the results of the previous approaches. The proposed model overcomes the published results on this dataset and obtains a new state-of-the-art with 87.7% as the PCK^h score. Also, our model achieves an increase of 1.4% and 0.2% over the closest approach [9] for the wrist and ankle parts respectively. This affirms the validity of the proposed method and the significance of encoding the hierarchical constraints into the model.

TABLE 3. PCK^h with $r = 0.5$ on LIP dataset.

Method	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	Mean
Hybrid Pose Machine	71.7	87.1	82.3	78.2	69.2	77.0	73.5	77.2
BUPTMM-POSE	90.4	87.3	81.9	78.8	68.5	75.3	75.8	80.2
Pyramid Stream Network	91.1	88.4	82.2	79.4	70.1	80.8	81.2	82.1
Nie *et al*. [28]	**94.9**	93.1	**89.9**	87.6	75.9	84.9	84.4	87.5
Chou *et al*. [9]	**94.9**	93.1	89.1	86.5	75.7	85.5	85.7	87.4
Proposed method	94.6	**93.2**	89.0	**87.9**	**77.5**	**85.9**	**85.9**	**87.7**

A visual comparison between our method and the approach of Chou *et al*. [9] is shown in Fig. 5-(c) on samples from the LIP dataset. The visualized samples show that the proposed method is able to generate better results in difficult conditions such as when the parts are fully occluded or there are large variations in the scale of the shown parts.

C. ABLATION STUDY

This study proposes using the two-stage of a parent-child hierarchy, with and without incorporating the adversarial branch. This assists in gaining insight on where the gains in the performance stem from. To achieve this, we design experiments, which study the two main metrics: adversarial training and hierarchical parent-child relationships while training the proposed method. This results in training four models on the same training datasets: (1) Two models for involving the

two-stage hierarchical representation in the trained models with and without using a discriminator and the adversarial loss. (2) The other two models are without the hierarchical representation. The number of the stacks in these experiments are fixed to two for all of the trained models.

Table 4 demonstrates the *PCK* of the entire body parts on 1000 samples of the testing images from the LSP dataset. Using the hierarchical relationships between the parents and children parts shows gains in the performance over when not encoding these relationships even when using the adversarial training. Also, it is noticeable that the combination of training the model achieves the best results, along with encoding the hierarchical relationships of the body parts and the adversarial training.

Other experiments of using Hard-Keypoints Mining (HKM) [42] were conducted on top of the proposed method to determine, whether selecting the keypoints with high losses to penalize the trained network would enhance the performance. The keypoints were selected based on their loss values and only the top n keypoints (in our case, $n = 8$) with high loss are used in the loss terms. As listed in Table 4, the results were degraded for most of the body parts, which indicates that using the HKM does not improve the performance as it considers a subset of the body parts.

TABLE 4. Total *PCK* with $r = 0.2$ on LSP dataset with number of stacks = 2.

Method	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	Total
No adversarial & no hierarchy	98.1	93.1	89.8	88.2	92.8	94.2	92.5	92.7
With adversarial & no hierarchy:	98.0	93.7	90.1	87.8	92.9	94.2	93.1	92.8
No adversarial & with hierarchy	98.4	94.0	91.5	88.1	93.6	94.2	92.6	93.2
With adversarial & with hierarchy	98.2	94.5	91.9	88.9	94.1	94.2	94.2	93.7
With adversarial & with hierarchy & with HKM [42]	98.0	94.2	90.7	88.6	93.1	93.8	93.2	93.1

V. CONCLUSION

In this paper, we have proposed a generative adversarial based network, which implicitly encodes the structure and hierarchy of body parts. The proposed network has shown the capability of learning the spatial relationships among the estimated body parts. It has been trained end-to-end and tested on three different benchmarks. The proposed network has shown the ability to estimate parts, which are largely deformable and highly occluded. The proposed method has overcome the performance of other approaches and achieved state-of-the-art results on some of the datasets and has been comparable on others. The proposed network can be extended to encode other association features between the adjacent body parts. Moreover, the proposed network can be adapted to handle multiple levels of hierarchy among the body parts.

ACKNOWLEDGMENT

The authors would like to thank the support of the NVIDIA Corporation with the donation of one NVIDIA TITAN Xp GPU for this research.

REFERENCES

[1] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2D human pose estimation: New benchmark and state of the art analysis," in *Proc. CVPR*, Jun. 2014, pp. 3686–3693.
[2] M. Andriluka, S. Roth, and B. Schiele, "Pictorial structures revisited: People detection and articulated pose estimation," in *Proc. CVPR*, Jun. 2009, pp. 1014–1021.
[3] V. Belagiannis and A. Zisserman, "Recurrent human pose estimation," in *Proc. FG*, May/Jun. 2017, pp. 468–475.
[4] D. Berthelot, T. Schumm, and L. Metz, "Began: Boundary equilibrium generative adversarial networks," Mar. 2017, *arXiv:1703.10717*. [Online]. Available: https://arxiv.org/abs/1703.10717
[5] A. Bulat and G. Tzimiropoulos, "Human pose estimation via convolutional part heatmap regression," in *Proc. ECCV*, 2016, pp. 717–732.
[6] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. CVPR*, Jul. 2017, pp. 7291–7299.
[7] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik, "Human pose estimation with iterative error feedback," in *Proc. CVPR*, Jun. 2016, pp. 4733–4742.
[8] Y. Chen, C. Shen, X.-S. Wei, L. Liu, and J. Yang, "Adversarial posenet: A structure-aware convolutional network for human pose estimation," in *Proc. ICCV*, Oct. 2017, pp. 1212–1221.
[9] C.-J. Chou, J.-T. Chien, and H.-T. Chen, "Self adversarial training for human pose estimation," Jul. 2017, *arXiv:1707.02439*. [Online]. Available: https://arxiv.org/abs/1707.02439
[10] X. Chu, W. Ouyang, H. Li, and X. Wang, "Structured feature learning for pose estimation," in *Proc. CVPR*, Jun. 2016, pp. 4715–4723.
[11] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang, "Multi-context attention for human pose estimation," in *Proc. CVPR*, Jul. 2017, pp. 1831–1840.
[12] R. Collobert, K. Kavukcuoglu, and C. Farabet, "Torch7: A matlab-like environment for machine learning," in *Proc. BigLearn, NIPS Workshop*, 2011, pp. 1–6.
[13] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. CVPR*, Jun. 2005, pp. 886–893.
[14] M. Eichner, V. Ferrari, and S. Zurich, "Better appearance models for pictorial structures," in *Proc. BMVC*, 2009, pp. 1–12.
[15] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *Proc. CVPR*, Jun. 2008, pp. 1–8.
[16] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester, "Cascade object detection with deformable part models," in *Proc. CVPR*, Jun. 2010, pp. 2241–2248.
[17] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient matching of pictorial structures," in *Proc. CVPR*, Jun. 2000, pp. 66–73.
[18] G. Gkioxari, A. Toshev, and N. Jaitly, "Chained predictions using convolutional neural networks," in *Proc. ECCV*, 2016, pp. 728–743.
[19] K. Gong, X. Liang, D. Zhang, X. Shen, and L. Lin, "Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing," in *Proc. CVPR*, Jul. 2017, pp. 932–940.
[20] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. NIPS*, 2014, pp. 1–9.
[21] P. Hu and D. Ramanan, "Bottom-up and top-down reasoning with hierarchical rectified gaussians," in *Proc. CVPR*, Jun. 2016, pp. 5600–5609.
[22] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, "DeeperCut: A deeper, stronger, and faster multi-person pose estimation model," in *Proc. ECCV*, 2016, pp. 34–50.
[23] S. Johnson and M. Everingham, "Clustered pose and nonlinear appearance models for human pose estimation," in *Proc. BMVC*, 2010, pp. 1–11.
[24] L. Ke, M.-C. Chang, H. Qi, and S. Lyu, "Multi-scale structure-aware network for human pose estimation," in *Proc. ECCV*, 2018, pp. 713–728.
[25] I. Lifshitz, E. Fetaya, and S. Ullman, "Human pose estimation using deep consensus voting," in *Proc. ECCV*, 2016, pp. 246–260.
[26] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. ECCV*, 2016, pp. 483–499.
[27] X. Nie, J. Feng, J. Xing, S. Xiao, and S. Yan, "Hierarchical contextual refinement networks for human pose estimation," *IEEE Trans. Image Process.*, vol. 28, no. 2, pp. 924–936, Feb. 2019.
[28] X. Nie, J. Feng, Y. Zuo, and S. Yan, "Human pose estimation with parsing induced learner," in *Proc. CVPR*, Jun. 2018, pp. 2100–2108.

[29] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele, "Strong appearance and expressive spatial models for human pose estimation," in *Proc. ICCV*, Dec. 2013, pp. 3487–3494.
[30] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. V. Gehler, and B. Schiele, "DeepCut: Joint subset partition and labeling for multi person pose estimation," in *Proc. CVPR*, Jun. 2016, pp. 4929–4937.
[31] U. Rafi, B. Leibe, J. Gall, and I. Kostrikov, "An efficient convolutional network for human pose estimation," in *Proc. BMVC*, 2016, pp. 1–11.
[32] X. Sun, J. Shang, S. Liang, and Y. Wei, "Compositional human pose regression," in *Proc. ICCV*, Oct. 2017, pp. 2602–2611.
[33] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler, "Efficient object localization using convolutional networks," in *Proc. CVPR*, Jun. 2015, pp. 648–656.
[34] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," in *Proc. NIPS*, 2014, pp. 1–9.
[35] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *Proc. CVPR*, Jun. 2014, pp. 1653–1660.
[36] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proc. CVPR*, Jun. 2016, pp. 4724–4732.
[37] W. Yang, S. Li, W. Ouyang, H. Li, and X. Wang, "Learning feature pyramids for human pose estimation," in *Proc. ICCV*, Oct. 2017, pp. 1281–1290.
[38] W. Yang, W. Ouyang, H. Li, and X. Wang, "End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation," in *Proc. CVPR*, Jun. 2016, pp. 3073–3082.
[39] Y. Yang and D. Ramanan, "Articulated pose estimation with flexible mixtures-of-parts," in *Proc. CVPR*, Jun. 2011, pp. 1385–1392.
[40] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. S. Torr, "Conditional random fields as recurrent neural networks," in *Proc. ICCV*, Dec. 2015, pp. 1529–1537.
[41] S. Honariand, P. Molchanov, S. Tyree, P. Vincent, C. Pal, and J. Kautz, "Improving landmark localization with semi-supervised learning," in *Proc. CVPR*, Jun. 2018, pp. 1546–1555.
[42] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, "Cascaded pyramid network for multi-person pose estimation," in *Proc. CVPR*, Jun. 2018, pp. 7103–7112.
[43] Z. Tang, X. Peng, S. Geng, L. Wu, S. Zhang, and D. Metaxas, "Quantized densely connected u-nets for efficient landmark localization," in *Proc. ECCV*, Sep. 2018, pp. 339–354.
[44] W. Tang, P. Yu, and Y. Wu, "Deeply learned compositional models for human pose estimation," in *Proc. ECCV*, 2018, pp. 190–206.

IBRAHIM RADWAN received the bachelor's and master's degrees in computer science from the Faculty of Computer and Information, Zagazig University, Egypt, in 2004 and 2009, respectively, and the Ph.D. degree in computer vision from the University of Canberra, in 2015. From 2014 to 2016, he was a Computer Vision Researcher at a leading automotive industry warehouse. He is currently a Research Fellow with The Australian National University (ANU), Canberra, Australia. His research interests include computer vision, machine learning, robotics, and artificial intelligence.

NOUR MOUSTAFA received the bachelor's and master's degrees in computer science from the Faculty of Computer and Information, Helwan University, Egypt, in 2009 and 2014, respectively, and the Ph.D. degree in cybersecurity from the University of New South Wales (UNSW), Canberra, Australia, in 2017. He was a Postdoctoral Fellow with UNSW, from June 2017 to December 2018. He is currently a Lecturer with SEIT, UNSW, and Helwan University, Egypt. His areas of interests include cybersecurity, in particular, network security, host- and network-intrusion detection systems, statistics, deep learning, and machine learning techniques. He is interested in designing and developing threat detection and forensic mechanisms to the Industry 4.0 technology for identifying malicious activities from cloud computing, fog computing, the IoT, and industrial control systems over virtual machines and physical systems.

BYRON KEATING is currently a Professor with the QUT Business School, Queensland University of Technology. He is also the Director of the Service Innovation Lab, a service-focused management and research consultancy based in Brisbane, Australia. His research interests are concerned with the role of emerging technologies in supporting the design and delivery of complex services. This interest began with his Ph.D. research at The University of Newcastle and continues today in the areas of platforms, location-based services, and artificial intelligence. His work has been published in journals, such as the Proceedings of the IEEE, the *European Journal of Information Systems*, the *Journal of Supply Chain Management*, and the *Journal of Service Management*.

KIM-KWANG RAYMOND CHOO (SM'15) received the Ph.D. degree in information security from the Queensland University of Technology, Australia, in 2006. He currently holds the Cloud Technology Endowed Professorship at The University of Texas at San Antonio. He is also a Fellow of the Australian Computer Society and the Co-Chair of the IEEE Multimedia Communications Technical Committee's Digital Rights Management for Multimedia Interest Group. In 2016, he was named the Cybersecurity Educator of the Year–APAC (Cybersecurity Excellence Awards are produced in cooperation with the Information Security Community on LinkedIn), and in 2015, he and his team won the Digital Forensics Research Challenge organized by Germany's University of Erlangen-Nuremberg. He is a recipient of the 2019 IEEE Technical Committee on Scalable Computing (TCSC) Award for Excellence in Scalable Computing (Middle Career Researcher), the 2018 UTSA College of Business Col. Jean Piccione and Lt. Col. Philip Piccione Endowed Research Award for Tenured Faculty, the Outstanding Associate Editor of 2018 for IEEE Access, the British Computer Society's 2019 Wilkes Award Runner-up, the 2019 *EURASIP Journal on Wireless Communications and Networking* (JWCN) Best Paper Award, the Korea Information Processing Society's *Journal of Information Processing Systems* (JIPS) Survey Paper Award (Gold) 2019, the IEEE Blockchain 2019 Outstanding Paper Award, the IEEE TrustCom 2018 Best Paper Award, the ESORICS 2015 Best Research Paper Award, the 2014 Highly Commended Award from the Australia New Zealand Policing Advisory Agency, the Fulbright Scholarship in 2009, the 2008 Australia Day Achievement Medallion, and the British Computer Society's Wilkes Award in 2008.

ROLAND GOECKE received the master's degree in computer science from the University of Rostock, Germany, in 1998, and the Ph.D. degree in computer science from The Australian National University, Canberra, ACT, Australia, in 2004. He is currently a Professor of affective computing with the University of Canberra, where he is also the Director of the Human-Centered Technology Research Centre and leads the Vision and Sensing Group. His research interests include affective computing, pattern recognition, computer vision, human–computer interaction, and multimodal signal processing.